10/088895
PCT/FR00/02640
FR00/2640

R E P U B L I Q U E  F R A N Ç A I S E

INPI
INSTITUT NATIONAL DE LA PROPRIETE INDUSTRIELLE



# BREVET D'INVENTION

4

## CERTIFICAT D'UTILITÉ - CERTIFICAT D'ADDITION

## COPIE OFFICIELLE

Le Directeur général de l'Institut national de la propriété industrielle certifie que le document ci-annexé est la copie certifiée conforme d'une demande de titre de propriété industrielle déposée à l'Institut.

Fait à Paris, le 04 OCT. 2000

Pour le Directeur général de l'Institut
national de la propriété industrielle
Le Chef du Département des brevets

Martine PLANCHE

**DOCUMENT DE PRIORITÉ**

**PRÉSENTÉ OU TRANSMIS CONFORMÉMENT À LA REGLE 17.1.a) OU b)**

INSTITUT NATIONAL DE LA PROPRIETE INDUSTRIELLE
SIEGE
26 bis, rue de Saint Petersbourg
75800 PARIS Cédex 08
Téléphone : 01 53 04 53 04
Télécopie : 01 42 93 59 30

INPI
INSTITUT NATIONAL DE LA PROPRIETE INDUSTRIELLE

# BREVET D'INVENTION, CERTIFICAT D'UTILITÉ

Code de la propriété intellectuelle-Livre VI

cerfa N° 55 -1328

## REQUÊTE EN DÉLIVRANCE

26 bis, rue de Saint Pétersbourg
75800 Paris Cedex 08
Téléphone : 01 53 04 53 04 Télécopie : 01 42 93 59 30

Confirmation d'un dépôt par télécopie ☐

Cet imprimé est à remplir à l'encre noire en lettres capitales

Réservé à l'INPI

DATE DE REMISE DES PIÈCES 24 SEPT 1999
N° D'ENREGISTREMENT NATIONAL 9911973
DÉPARTEMENT DE DÉPÔT 75 INPI PARIS
DATE DE DÉPÔT 24 SEP. 1999

1 NOM ET ADRESSE DU DEMANDEUR OU DU MANDATAIRE À QUI LA CORRESPONDANCE DOIT ÊTRE ADRESSÉE

CABINET LAVOIX
2 Place d'Estienne d'Orves
75441 PARIS CEDEX 09

n°du pouvoir permanent références du correspondant téléphone
BFF 99/0302 53-20-14-20

2 DEMANDE Nature du titre de propriété industrielle

☒ brevet d'invention ☐ demande divisionnaire ☐ demande initiale
☐ certificat d'utilité ☐ transformation d'une demande de brevet européen ☐ brevet d'invention ☐ certificat d'utilité n° date

Établissement du rapport de recherche ☐ différé ☒ immédiat

Le demandeur, personne physique, requiert le paiement échelonné de la redevance ☐ oui ☐ non

Titre de l'invention (200 caractères maximum)

**Procédé de classification thématique de documents, module de classification thématique et moteur de recherche incorporant un tel module.**

3 DEMANDEUR (S) n° SIREN code APE-NAF

Nom et prénoms (souligner le nom patronymique) ou dénomination

FRANCE TELECOM

Forme juridique

Nationalité (s) Française

Adresse (s) complète (s) Pays

6, Place d'Alleray 75015 PARIS FR

En cas d'insuffisance de place, poursuivre sur papier libre ☐

4 INVENTEUR (S) Les inventeurs sont les demandeurs ☐ oui ☒ non Si la réponse est non, fournir une désignation séparée

5 RÉDUCTION DU TAUX DES REDEVANCES ☐ requise pour la 1ère fois ☐ requise antérieurement au dépôt ; joindre copie de la décision d'admission

6 DÉCLARATION DE PRIORITÉ OU REQUÊTE DU BÉNÉFICE DE LA DATE DE DÉPÔT D'UNE DEMANDE ANTÉRIEURE

| pays d'origine | numéro | date de dépôt | nature de la demande |
|---|---|---|---|
| | | | |

7 DIVISIONS antérieures à la présente demande n° date n° date

8 SIGNATURE DU DEMANDEUR OU DU MANDATAIRE (nom et qualité du signataire)

CABINET LAVOIX
H. OBOLENSKY n° 92-1186

SIGNATURE DU PRÉPOSÉ À LA RÉCEPTION

SIGNATURE APRÈS ENREGISTREMENT DE LA DEMANDE À L'INPI

INPI
INSTITUT NATIONAL DE LA PROPRIETE INDUSTRIELLE

# BREVET D'INVENTION, CERTIFICAT D'UTILITE

## DÉSIGNATION DE L'INVENTEUR
(si le demandeur n'est pas l'inventeur ou l'unique inventeur)


DEPARTEMENT DES BREVETS

26bis, rue de Saint-Pétersbourg
75800 Paris Cédex 08
Tél. : 01 53 04 53 04 - Télécopie : 01 42 93 59 30


N° D'ENREGISTREMENT NATIONAL

9911573

TITRE DE L'INVENTION : Procédé de classification thématique de documents, module de classification thématique et moteur de recherche incorporant un tel module.

LE(S) SOUSSIGNÉ(S)

FRANCE TELECOM
6, Place d'Alleray 75015 PARIS FRANCE

DÉSIGNE(NT) EN TANT QU'INVENTEUR(S) (indiquer nom, prénoms, adresse et souligner le nom patronymique) :

BIETTRON Laurent
Kerauzern
22300 PLOUBEZRE FRANCE

PALLU Frédéric
Keravel
22560 TREBEURDEN FRANCE

TRICOT Sylvie
14, Hent Lann
22300 TREDREZ FRANCE

NOTA : A titre exceptionnel, le nom de l'inventeur peut être suivi de celui de la société à laquelle il appartient (société d'appartenance) lorsque celle-ci est différente de la société déposante ou titulaire.

Date et signature (s) du (des) demandeur (s) ou du mandataire

Paris, le 24 septembre 1999

CABINET LAVOIX
M. OBOLENSKY n° 92.1186

BA 113/140897

## DOCUMENT COMPORTANT DES MODIFICATIONS

| PAGE(S) DE LA DESCRIPTION OU DES REVENDICATIONS OU PLANCHE(S) DE DESSIN | | | R.M.* | DATE DE LA CORRESPONDANCE | TAMPON DATEUR DU CORRECTEUR |
|---|---|---|---|---|---|
| Modifiée(s) | Supprimée(s) | Ajoutée(s) | | | |
| 12-13-14-15 | | | α | 19/5/00 | 30 MAI 2000 - B B |
| | | | | | 16 AOUT 2000 - B B |
| Page 13 | | | | 10/8/00 | |
| | | | | | |
| | | | | | |
| | | | | | |
| | | | | | |
| | | | | | |
| | | | | | |

Un changement apporté à la rédaction des revendications d'origine, sauf si celui-ci découle des dispositions de l'article R.612-36 du code de la Propriété Intellectuelle, est signalé par la mention «R.M.» (revendications modifées).

La présente invention se rapporte à un procédé de classification thématique de documents, destiné, en particulier, à la constitution ou la mise à jour de bases de données thématiques, en particulier pour moteur de recherche.

Elle se rapporte également à un module de classification thématique de documents et à un moteur de recherche équipé d'un tel module de classification thématique.

On connaît, à ce jour, principalement deux outils informatiques permettant de rechercher des documents sur un réseau informatique, comme par exemple, le réseau Internet.

Ces outils sont le moteur de recherche et le guide.

Un moteur de recherche est un outil permettant d'extraire d'une information, principalement textuelle, les mots ou termes qui la représentent le mieux et de les stocker dans des bases de données, également connues sous l'appellation "base d'index".

De telles bases d'index sont généralement mises à jour relativement fréquemment.

En réponse à une requête formulée par un utilisateur, ce même outil parcourt les bases d'index afin d'identifier les termes les plus pertinents par rapport à ceux de la requête, puis de trier les informations à fournir en retour.

L'autre technique de recherche de documents sur un réseau informatique consiste à utiliser un guide. Cet outil propose des recherches par catégories, les pages de documents étant classées manuellement par des documentalistes.

Ces types d'outil présentent un certain nombre d'inconvénients.

Tout d'abord, les moteurs de recherche ne proposent pas de classement de pages de document par catégories. En effet, les pages fournies en réponse à une requête ne sont pas typées. Ainsi, des requêtes ambiguës peuvent

donner lieu à des réponses très diverses, ressenties comme du bruit par l'utilisateur.

Les guides, au contraire, permettent de fournir à un utilisateur des réponses typées, c'est à dire portant sur le ou les mêmes thèmes que la requête.

Cependant, le classement manuel des pages de document implique de forts coûts de création et de mise à jour et ne permet l'indexation que d'un nombre limité de pages. Par conséquent, certaines requêtes n'obtiennent pas de réponse.

Le but de l'invention est de palier les inconvénients des moteurs de recherche et des guides.

Elle a donc pour objet un procédé de classification thématique de documents, notamment, pour la constitution ou la mise à jour de bases de données thématiques pour moteur de recherche, caractérisé en ce qu'il comporte les étapes suivantes :

- on sélectionne un échantillon de documents représentatifs de chaque thème ;
- on identifie, dans les documents sélectionnés, des éléments caractéristiques de chaque thème ;
- on affecte, à chaque élément identifié, un coefficient représentatif de la pertinence de cet élément vis à vis du thème correspondant ; et
- pour chaque document à classifier, on identifie lesdits éléments caractéristiques de chaque thème qu'il contient et, pour chaque thème qui leur correspond, on calcule, à partir du coefficient affecté à ces éléments, la valeur d'une caractéristique représentative de la pertinence du thème pour ce document, pour décider si ce document porte ou non sur ce thème.

On classe ainsi les documents récupérés sur un réseau informatique en fonction des thèmes qui y sont abordés et ce, de façon automatique.

Le procédé de classification selon l'invention peut en outre comporter une ou plusieurs des caractéristiques suivantes, prises isolément ou selon toutes les combinaisons techniquement possibles :

- l'étape d'affectation dudit coefficient à chaque élément identifié comprend les étapes suivantes, pour chaque thème :

. calcul de la fréquence de l'élément dans les documents sélectionnés portant sur ce thème,

. calcul de la fréquence de l'élément dans les documents sélectionnés ne portant pas sur ce thème, et

. calcul du rapport entre les fréquences calculées.

- il comporte en outre une étape de tri des thèmes selon une arborescence de thèmes et par ordre décroissant des coefficients,

- l'étape de calcul de la caractéristique représentative de la pertinence du thème d'un document à classifier comprend les étapes suivantes pour chaque thème :

. on lit la valeur du rapport desdites fréquences de chaque élément représentatif du thème extrait du document,

. on multiplie les valeurs lues et

. on affecte le résultat de cette multiplication à la valeur de ladite caractéristique.

- l'on décide que le document porte sur un thème si la valeur de ladite caractéristique représentative de la pertinence du thème pour ce document est supérieure à une valeur de seuil,

- la valeur de seuil est élaborée, pour chaque thème, à partir desdits rapports de fréquence, selon la relation suivante :

$$score\_seuil_{thème} = (R_{moy})^{nthème}$$

dans laquelle :

$score\_seuil_{thème}$ désigne la valeur de seuil

$R_{moy}$ représente la valeur moyenne des rapports de fréquences R des éléments du thème et,

nthème désigne un nombre prédéterminé.

- selon une variante, la valeur de seuil est réglée manuellement.

- les étapes d'identification des éléments caractéristiques de chaque thème contenu dans un document sont réalisées au moyen d'une table de hachage.

- on calcule, pour chaque élément de vocabulaire d'une requête formulée par un utilisateur, des coefficients caractéristiques de l'élément par rapport à chaque thème connu et l'on associe à chaque élément les coefficients et les thèmes correspondants, de sorte que lesdits coefficients atteignent une valeur minimale.

Lors de la recherche des entrées d'index, c'est à dire au cours de la recherche des documents correspondants à la requête, il est ainsi possible d'accéder directement aux thèmes liés à chaque élément et aux coefficients correspondants que l'on combine par multiplication afin de déterminer un classement des thèmes liés à la requête entière.

L'invention a également pour objet un module de classification thématique de documents, notamment pour moteur de recherche, caractérisé en ce qu'il comporte des moyens de comparaison d'éléments extraits de chaque document avec des éléments caractéristiques de différents thèmes affectés chacun d'un coefficient représentatif de la pertinence de cet élément pour un thème correspondant et des moyens de calcul de la valeur d'au moins une caractéristique représentative de la pertinence d'un thème pour ce document, à partir des coefficients desdits éléments caractéristiques qu'il contient pour décider si ce document porte ou non sur ce thème.

Un autre objet de l'invention est un moteur de recherche de documents sur un réseau informatique, comprenant un module d'indexation pour la création et la mise à jour de bases de données thématiques, à partir de documents récupérés sur le réseau informatique, et un module d'interrogation des bases de données adaptées pour fournir des références de documents correspondant à une requête reçue en entrée, caractérisé en ce qu'il comporte en outre un module de classification thématique tel que définit ci-dessus, associé au module d'indexation.

D'autres caractéristiques et avantages ressortiront de la description suivante, donnée uniquement à titre d'exemple, et faite en référence aux dessins annexés sur lesquels :

- la Fig. 1 est un organigramme montrant les principales phases de fonctionnement d'un module de classification thématique de documents selon l'invention, pour moteur de recherche ;
- la Fig. 2 est un organigramme illustrant la méthode de calcul des éléments caractéristiques de thèmes ; et
- la Fig. 3 est un organigramme montrant la méthode de calcul des thèmes d'un document.

Sur la Fig. 1, on a représenté les principales phases du procédé de classification thématique de documents selon l'invention.

Il est destiné à permettre le classement de documents récupérés sur un réseau informatique, en fonction de thèmes qui y sont abordés. Par exemple, il peut être mis en oeuvre au sein d'un moteur de recherche.

Dans ce cas, il intervient dès le processus d'indexation, mais également au cours du traitement d'une requête formulée par un utilisateur, pour permettre de déterminer tous les thèmes abordés dans cette requête.

On conçoit toutefois que d'autres applications peuvent être envisagées. Par exemple, ce procédé peut être mis en oeuvre au niveau d'un point d'accès d'un réseau de postes utilisateurs à un réseau Internet, afin de déterminer la nature des pages Web récupérées par les utilisateurs et interdire ou autoriser, par filtrage des requêtes, certains thèmes, par exemple, contraires à l'ordre public et aux bonnes moeurs, ou encore calculer des statistiques sur les centres d'intérêt des utilisateurs.

Pour procéder à cette classification, le procédé comporte deux phases distinctes, à savoir une première phase préalable d'acquisition du vocabulaire thématique de corpus de documents et d'affectation, à chaque mot du

vocabulaire, d'une valeur de seuil à partir de laquelle on décide qu'un document, contenant ce mot, porte sur le thème correspondant, ainsi qu'une deuxième phase de classification proprement dite, au cours de laquelle un document récupéré sur le réseau est automatiquement classifié en fonction des éléments caractéristiques qu'il contient.

Par exemple cette deuxième phase intervient périodiquement, seuls des documents nouvellement créés ou modifiés étant classifiés.

La description de la première phase d'acquisition du vocabulaire thématique va maintenant être en référence aux Figs. 1 à 3.

Comme on le voit sur la Fig. 1, cette phase débute par une étape 10 de sélection manuelle, à partir d'un ensemble 12 d'échantillons (ou corpus) de documents représentatifs de chacun des thèmes A à Z utilisés pour classer les documents au cours de la deuxième phase.

Ainsi, à l'issu de cette étape 10 de sélection manuelle, on dispose d'un ensemble de corpus de documents, tels que 14, portant chacun sur un thème (thème A,... thème Z). Bien entendu l'étape de sélection peut également être effectuée par tout moyen autre que manuel.

Au cours de cette étape 10 de sélection, on crée également un corpus 16 de documents ne portant sur aucun des thèmes A à Z et on définit une nomenclature 18 des thèmes A à Z, c'est à dire la liste de ces thèmes associés à des sous-thèmes s'y rapportant.

Lors de l'étape 20 suivante, ces éléments sont présentés en entrée d'un module de classification thématique en vue d'extraire de chaque document les éléments caractéristiques de chaque thème et de les affecter chacun d'un coefficient représentatif de leur pertinence vis à vis d'un thème correspondant.

Par exemple ce module de classification thématique se présente sous la forme d'un module spécifique d'un moteur de recherche, associé à un module d'indexation

réalisant la création ou la mise à jour des bases de données thématiques.

Il peut également être agencé sous la forme d'un module spécifique prévu au niveau d'un point d'accès à un réseau informatique, en particulier à un réseau Internet.

Ce module comprend les moyens logiciels appropriés pour réaliser l'extraction des éléments caractéristiques de chaque thème et pour les affecter d'un coefficient représentatif de leur pertinence vis à vis de différents thèmes, comme cela va être décrit en détail par la suite.

Au cours de cette étape 20, le module de classification extrait, de chaque document sélectionné, les éléments caractéristiques de chaque thème.

Cette extraction s'effectue en utilisant un outil informatique de type classique. Il ne sera donc pas décrit par la suite.

On dispose à l'issu de cette étape 20, de listes d'éléments caractéristiques des thèmes A à Z, telles que 22.

En référence à la Fig. 2, cette procédure d'identification du vocabulaire caractéristique de chaque thème s'effectue successivement pour chaque élément extrait des documents de chacun des corpus 14 et 16.

Au cours d'une première étape 24, on vide un tableau regroupant l'ensemble des thèmes candidats, c'est à dire les thèmes susceptibles de correspondre à l'élément extrait.

Lors de l'étape 26 suivante, on procède, pour chaque thème, à un calcul d'un coefficient R représentatif de la pertinence de cet élément vis à vis de ce thème.

Pour procéder à ce calcul, on calcule tout d'abord la fréquence p de l'élément dans les documents portant sur ce thème, ainsi que la fréquence q de cet élément dans les documents ne portant pas sur ce thème.

On procède ensuite au calcul du coefficient R, constitué par le rapport entre ces fréquences p et q.

Lors de l'étape 28 suivante, on vérifie si les caractéristiques p, q et R se situent à l'intérieur de limites prédéterminées.

Si tel n'est pas le cas, on procède au traitement de l'élément suivant.

Si tel est le cas, on ajoute le thème dans le tableau des thèmes candidats avec un score égal au coefficient R (étape 30).

S'il reste des éléments à traiter (étape 32), la procédure retourne à l'étape 24 précédente.

Dans le cas contraire, cette procédure s'achève.

On notera que, de préférence, après remplissage du tableau des thèmes candidats, celui-ci est trié par ordre décroissant des scores R. On notera également que pour tout thème candidat, jusqu'à un nombre maximum voulu, on ajoute un nouvel élément récupéré dans la liste des éléments caractéristiques de ce thème, en se limitant à un nombre maximum voulu des n meilleurs éléments par thème choisi en fonction de leur score R.

En se référant à nouveau à la Fig. 1, lors de l'étape 34 suivante, le module de classification thématique procède à un calcul automatique, au moyen d'un algorithme approprié, d'une valeur de seuil correspondant à un seuil minimum à atteindre pour déterminer si un document comprenant un élément caractéristique d'un thème porte ou non sur ce thème.

Pour procéder à ce calcul, le module de classification procède tout d'abord à un calcul de la valeur moyenne $R_{moy}$ des rapports R des éléments caractéristiques de chaque thème (étape 36).

Il procède ensuite au calcul de la valeur de seuil $score\_seuil_{thème}$, selon la relation suivante :

$$score\_seuil_{thème} = (R_{moy})nthème$$

dans laquelle nthème désigne un nombre prédéterminé choisi par exemple égal à 5 pour la plupart des thèmes.

On voit alors sur la Fig. 1, qu'à l'issu de ce calcul automatique des scores à atteindre, on dispose de listes, telles que 40, d'éléments caractéristiques de

chaque thème A à Z, affectés chacun d'un score à atteindre, c'est à dire d'une valeur de seuil à partir de laquelle on considère qu'un document porte sur ce thème.

Après cette phase d'acquisition du vocabulaire thématique, réalisée à partir de corpus de documents représentatifs de thèmes, la deuxième phase de classification thématique proprement dite peut être effectuée, dans le but de constituer des bases de données thématiques, désignées par la référence numérique générale 42, à partir de documents collectés automatiquement sur le réseau informatique par des robots, tels que 44.

Ces documents sont présentés en entrée du module de classification thématique, qui reçoit également une indication de la nomenclature 18 des thèmes, ainsi que les éléments disponibles à l'issu de l'étape 34 mentionnée précédemment. Ce module procède à un calcul automatique des thèmes sur lesquels porte le document (étape 46).

Pour ce faire, il comporte tous les moyens logiciels appropriés pour réaliser les opérations mentionnées ci-dessous.

En référence à la Fig. 3, au cours d'une première étape 48 de cette procédure, le module d'indexation extrait de chaque document 50 récupéré par les robots 44, les éléments caractéristiques de thèmes qu'il contient.

Cette étape s'effectue, par exemple, en utilisant une table de hachage, pour rechercher rapidement dans les listes d'éléments caractéristiques les éléments contenus dans chaque document.

Après extraction de ces éléments on identifie, parmi ceux-ci, les éléments caractéristiques de thèmes contenus dans les listes 40.

Pour chaque élément identifié, le module de classification procède ensuite à un calcul d'une valeur caractéristique représentative de la pertinence de chaque thème pour ce document, à partir du coefficient affecté à cet élément.

Pour ce faire, lors de l'étape 52 suivante, une variable "score_thème" , représentative du score du document dans un thème donné est positionnée à 1, et ce pour tous les thèmes.

Ensuite, pour tout élément du document, et pour chaque thème de l'arborescence des thèmes, si l'élément se situe parmi la liste des éléments caractéristiques du thème, on lit le score R, c'est à dire la valeur du rapport des fréquences pour chaque élément et on multiplie les valeurs lues du score R pour chacun de ces éléments.

Le résultat de cette multiplication est ensuite affecté à la valeur de la caractéristique score _ thème(étape 54).

On décide alors que les thèmes reconnus dans le document 50 sont ceux dont la caractéristique score _ thème atteint ou dépasse le score à atteindre pour ces thèmes (étape 56).

On dispose alors, à l'issu de cette procédure, de l'ensemble 57 des thèmes sur le ou lesquels porte le document 50 récupéré.

On conçoit donc que cette procédure de calcul automatique des thèmes des documents récupérés par les robots 44 permet au module d'indexation d'un moteur de recherche de classer ces documents en fonction des thèmes abordés et de constituer les bases 42 de données thématiques.

Une telle procédure de calcul automatique de thème de documents peut également être utilisée pour déterminer les thèmes abordés dans une requête formulée par un utilisateur.

Pour ce faire, à partir de cette requête, pour chacun des éléments du vocabulaire d'interrogation utilisés dans la requête, on calcule les coefficients caractéristiques de cet élément par rapport à chacun des thèmes connus et l'on associe à chacun de ces éléments les coefficients et thèmes de telle manière que les coefficients atteignent une valeur minimale.

Lors de la recherche des entrées d'index correspondant aux éléments d'une requête, c'est à dire pour le calcul des résultats, on accède ainsi directement au thème lié aux éléments ainsi qu'à leur coefficient, que l'on combine par multiplication, selon la même procédure que celle décrite plus haut, afin de déterminer un classement des thèmes liés à la requête entière.

On conçoit donc que cette procédure permet de proposer à un utilisateur de préciser sa requête, par exemple, lorsque celle-ci est formulée de façon vague.

On conçoit également que cette procédure, qui permet d'identifier les thèmes contenus dans une requête, rend possible d'effectuer une surveillance des requêtes utilisateurs afin d'établir des calculs statistiques permettant de définir des profils d'utilisateurs en fonction des requêtes.

On saisira alors que l'invention qui vient d'être décrite peut être utilisée pour la recherche de thèmes contenus dans des pages récupérées sur un réseau informatique, pour la détermination de thèmes contenus dans une requête formulée par un utilisateur et, à partir de cette détermination, pour le filtrage des requêtes et également des pages récupérées, afin d'interdire la formulation de requête ou la récupération de pages portant sur des thèmes prédéterminés interdits, et pour l'élaboration des profils d'utilisateurs.

On notera cependant que dans le cas de la détermination des thèmes contenus dans une requête, cette dernière est considérée comme constituant un document présenté en entrée du module de classification thématique selon l'invention.

L'invention n'est pas limitée au mode de réalisation envisagée.

En effet, il est également possible, en variante, de régler manuellement la valeur de seuil à partir de laquelle on décide qu'un document porte ou non sur un thème donné.

**REVENDICATIONS**

1. Procédé de classification thématique de documents, notamment pour la constitution ou la mise à jour de bases de données thématiques pour moteur de recherche, caractérisé en ce qu'il comporte les étapes suivantes :

- on sélectionne un échantillon de documents représentatifs de chaque thème ;
- on identifie, dans les documents sélectionnés, des éléments caractéristiques de chaque thème ;
- on affecte, à chaque élément identifié, un coefficient (R) représentatif de la pertinence de cet élément vis à vis du thème correspondant ; et
- pour chaque document (50) à classifier, on identifie lesdits éléments caractéristiques de chaque thème qu'il contient et, pour chaque thème qui leur correspond, on calcule, à partir du coefficient affecté à ces éléments, la valeur d'une caractéristique représentative de la pertinence du thème pour ce document (50), pour décider si ce document porte ou non sur ce thème.

2. Procédé selon la revendication 1, caractérisé en ce que l'étape d'affectation dudit coefficient à chaque élément identifié comprend les étapes suivantes, pour chaque thème :

- calcul de la fréquence de l'élément dans les documents sélectionnés portant sur ce thème,
- calcul de la fréquence de l'élément dans les documents sélectionné ne portant pas sur ce thème, et
- calcul du rapport entre les fréquences calculées.

3. Procédé selon la revendication 2, caractérisé en ce qu'il comporte en outre une étape de tri des thèmes selon une arborescence de thèmes et par ordre décroissant des coefficients.

4. Procédé selon l'une des revendications 2 et 4, caractérisé en ce que l'étape de calcul de la caractéristique représentative de la pertinence du thème

d'un document à classifier comprend les étapes suivantes, pour chaque thème :

- on lit la valeur du rapport (R) desdites fréquences de chaque élément représentatif du thème extrait du document,
- on multiplie les valeurs lues, et
- on affecte le résultat de cette multiplication à la valeur de ladite caractéristique.

5. Procédé selon l'une quelconque des revendications 1 à 4, caractérisé en ce que l'on décide que le document porte sur un thème si la valeur de ladite caractéristique représentative de la pertinence du thème pour ce document est supérieure à une valeur de seuil.

6. Procédé selon la revendication 5, caractérisé en ce que la valeur de seuil est élaborée, pour chaque thème, à partir desdits rapports de fréquence, selon la relation suivante :

$$score\_seuil_{thème} = (R_{moy})nthème$$

dans laquelle :

$score\_seuil_{thème}$ désigne la valeur de seuil

$R_{moy}$ représente la valeur moyenne des rapports de fréquences R des éléments du thème et,

nthème désigne un nombre prédéterminé.

7. Procédé selon la revendication 5, caractérisé en ce que la valeur de seuil est réglée manuellement.

8. Procédé selon l'une quelconque des revendications 1 à 7, caractérisé en ce que les étapes d'identification des éléments caractéristiques de chaque thème contenu dans un document (50) sont réalisées au moyen d'une table de hachage.

9. Procédé selon l'une quelconque des revendications 1 à 8, caractérisé en ce que l'on calcule, pour chaque élément de vocabulaire d'une requête formulée par l'utilisateur, des coefficients caractéristiques de l'élément par rapport à chaque thème connu et l'on associe à chaque élément les coefficients et les thèmes correspondant, de sorte que lesdits coefficients atteignent une valeur minimale.

10. Module de classification thématique de documents (50), notamment pour moteur de recherche, caractérisé en ce qu'il comporte des moyens de comparaison d'éléments extraits de chaque document avec des éléments caractéristiques de différents thèmes, affectés chacun d'un coefficient (R) représentatif de la pertinence de cet élément pour un thème correspondant, et des moyens de calcul de la valeur d'au moins une caractéristique représentative de la pertinence d'un thème pour ce document, à partir des coefficients desdits éléments caractéristiques qu'il contient, pour décider si ce document (50) porte ou non sur ce thème.

11. Utilisation d'un module de classification thématique de documents selon la revendication 10 pour la recherche de thèmes contenus dans des pages récupérées sur un réseau informatique.

12. Utilisation d'un module de classification thématique de documents selon la revendication 10 pour la détermination de thèmes contenus dans une requête formulée par un utilisateur.

13. Utilisation d'un module de classification thématique de documents selon la revendication 10 pour la détermination de thèmes contenus dans des pages récupérées sur un réseau informatique ou dans une requête formulée par un utilisateur et le filtrage des documents récupérés pour interdire la consultation de pages portant sur un ou des thèmes prédéterminés.

14. Utilisation d'un module de classification thématique de documents selon la revendication 10 pour la détermination de thèmes contenus dans une requête formulée par un utilisateur et l'élaboration de profils d'utilisateurs à partir des thèmes sur lesquels porte la requête.

15. Moteur de recherche de documents sur un réseau informatique, comprenant un module d'indexation pour la création et la mise à jour de bases de données thématiques, à partir de documents récupérés sur le réseau informatique, et un module d'interrogation des bases de données thématiques adaptées pour fournir des

références de documents correspondant à une requête reçue en entrée, caractérisé en ce qu'il comporte en outre un module de classification thématique selon la revendication 10, associé au module d'indexation.

Feuille rectifiée

**REVENDICATIONS**

1. Procédé de classification thématique de documents, notamment pour la constitution ou la mise à jour de bases de données thématiques pour moteur de recherche, caractérisé en ce qu'il comporte les étapes suivantes :

- on sélectionne un échantillon de documents représentatifs de chaque thème ;

- on identifie, dans les documents sélectionnés, des éléments caractéristiques de chaque thème ;

- on affecte, à chaque élément identifié, un coefficient (R) représentatif de la pertinence de cet élément vis à vis du thème correspondant ;

- pour chaque document (50) à classifier, on identifie lesdits éléments caractéristiques de chaque thème qu'il contient et, pour chaque thème qui leur correspond, on calcule, à partir du coefficient affecté à ces éléments, la valeur d'une caractéristique représentative de la pertinence du thème pour ce document (50), pour décider si ce document porte ou non sur ce thème, lesdites étapes d'identification et de calcul étant réalisées automatiquement pour chaque document récupéré sur un réseau informatique ;

- on classe les documents récupérés en fonction des thèmes qui y sont abordés ; et

- l'on stocke les documents classés par thèmes dans des bases de données interrogeables à partir de thèmes contenus dans une requête

2. Procédé selon la revendication 1, caractérisé en ce que l'étape d'affectation dudit coefficient à chaque élément identifié comprend les étapes suivantes, pour chaque thème :

- calcul de la fréquence de l'élément dans les documents sélectionnés portant sur ce thème,

- calcul de la fréquence de l'élément dans les documents sélectionné ne portant pas sur ce thème, et

- calcul du rapport entre les fréquences calculées.

3. Procédé selon la revendication 2, caractérisé en ce qu'il comporte en outre une étape de tri des thèmes selon une arborescence de thèmes et par ordre décroissant des coefficients.

4. Procédé selon l'une des revendications 2 et 4, caractérisé en ce que l'étape de calcul de la caractéristique représentative de la pertinence du thème d'un document à classifier comprend les étapes suivantes, pour chaque thème :

- on lit la valeur du rapport (R) desdites fréquences de chaque élément représentatif du thème extrait du document,
- on multiplie les valeurs lues, et
- on affecte le résultat de cette multiplication à la valeur de ladite caractéristique.

5. Procédé selon l'une quelconque des revendications 1 à 4, caractérisé en ce que l'on décide que le document porte sur un thème si la valeur de ladite caractéristique représentative de la pertinence du thème pour ce document est supérieure à une valeur de seuil.

6. Procédé selon la revendication 5, caractérisé en ce que la valeur de seuil est élaborée, pour chaque thème, à partir desdits rapports de fréquence, selon la relation suivante :

$$\text{score} - \text{seuil}_{\text{thème}} = (R_{moy})\text{nthème}$$

dans laquelle :

score - $\text{seuil}_{\text{thème}}$ désigne la valeur de seuil

$R_{moy}$ représente la valeur moyenne des rapports de fréquences R des éléments du thème et,

nthème désigne un nombre prédéterminé.

7. Procédé selon la revendication 5, caractérisé en ce que la valeur de seuil est réglée manuellement.

8. Procédé selon l'une quelconque des revendications 1 à 7, caractérisé en ce que les étapes d'identification des éléments caractéristiques de chaque thème contenu dans un document (50) sont réalisées au moyen d'une table de hachage.

Feuille rectifiée

9. Procédé selon l'une quelconque des revendications 1 à 8, caractérisé en ce que l'on calcule, pour chaque élément de vocabulaire d'une requête formulée par l'utilisateur, des coefficients caractéristiques de l'élément par rapport à chaque thème connu et l'on associe à chaque élément les coefficients et les thèmes correspondant, de sorte que lesdits coefficients atteignent une valeur minimale.

10. Module de classification thématique de documents (50), notamment pour moteur de recherche, caractérisé en ce qu'il comporte une unité centrale de traitement comprenant des moyens de comparaison d'éléments extraits de chaque document avec des éléments caractéristiques de différents thèmes, affectés chacun d'un coefficient (R) représentatif de la pertinence de cet élément pour un thème correspondant, et des moyens de calcul de la valeur d'au moins une caractéristique représentative de la pertinence d'un thème pour ce document, à partir des coefficients desdits éléments caractéristiques qu'il contient, pour décider si ce document (50) porte ou non sur ce thème, ladite unité centrale étant raccordée à des moyens de stockage de documents classés par thèmes, interrogeables à partir de thèmes contenus dans une requête.

11. Utilisation d'un module de classification thématique de documents selon la revendication 10 pour la détermination de thèmes contenus dans une requête formulée par un utilisateur.

12. Utilisation d'un module de classification thématique de documents selon la revendication 10 pour la détermination de thèmes contenus dans des pages récupérées sur un réseau informatique ou dans une requête formulée par un utilisateur et le filtrage des documents récupérés pour interdire la consultation de pages portant sur un ou des thèmes prédéterminés.

13. Utilisation d'un module de classification thématique de documents selon la revendication 10 pour la détermination de thèmes contenus dans une requête formulée

par un utilisateur et l'élaboration de profils d'utilisateurs à partir des thèmes sur lesquels porte la requête.

14. Moteur de recherche de documents sur un réseau informatique, comprenant un module d'indexation pour la création et la mise à jour de bases de données thématiques, à partir de documents récupérés sur le réseau informatique, et un module d'interrogation des bases de données thématiques adaptées pour fournir des références de documents correspondant à une requête reçue en entrée, caractérisé en ce qu'il comporte en outre un module de classification thématique selon la revendication 10, associé au module d'indexation.

3. Procédé selon la revendication 2, caractérisé en ce qu'il comporte en outre une étape de tri des thèmes selon une arborescence de thèmes et par ordre décroissant des coefficients.

4. Procédé selon la revendication 2 ou 3, caractérisé en ce que l'étape de calcul de la caractéristique représentative de la pertinence du thème d'un document à classifier comprend les étapes suivantes, pour chaque thème :

- on lit la valeur du rapport (R) desdites fréquences de chaque élément représentatif du thème extrait du document,
- on multiplie les valeurs lues, et
- on affecte le résultat de cette multiplication à la valeur de ladite caractéristique.

5. Procédé selon l'une quelconque des revendications 1 à 4, caractérisé en ce que l'on décide que le document porte sur un thème si la valeur de ladite caractéristique représentative de la pertinence du thème pour ce document est supérieure à une valeur de seuil.

6. Procédé selon la revendication 5, caractérisé en ce que la valeur de seuil est élaborée, pour chaque thème, à partir desdits rapports de fréquence, selon la relation suivante :

$$\text{score} - \text{seuil}_{\text{thème}} = (R_{moy})^{\text{nthème}}$$

dans laquelle :

score - $\text{seuil}_{\text{thème}}$ désigne la valeur de seuil

$R_{moy}$ représente la valeur moyenne des rapports de fréquences R des éléments du thème et,

nthème désigne un nombre prédéterminé.

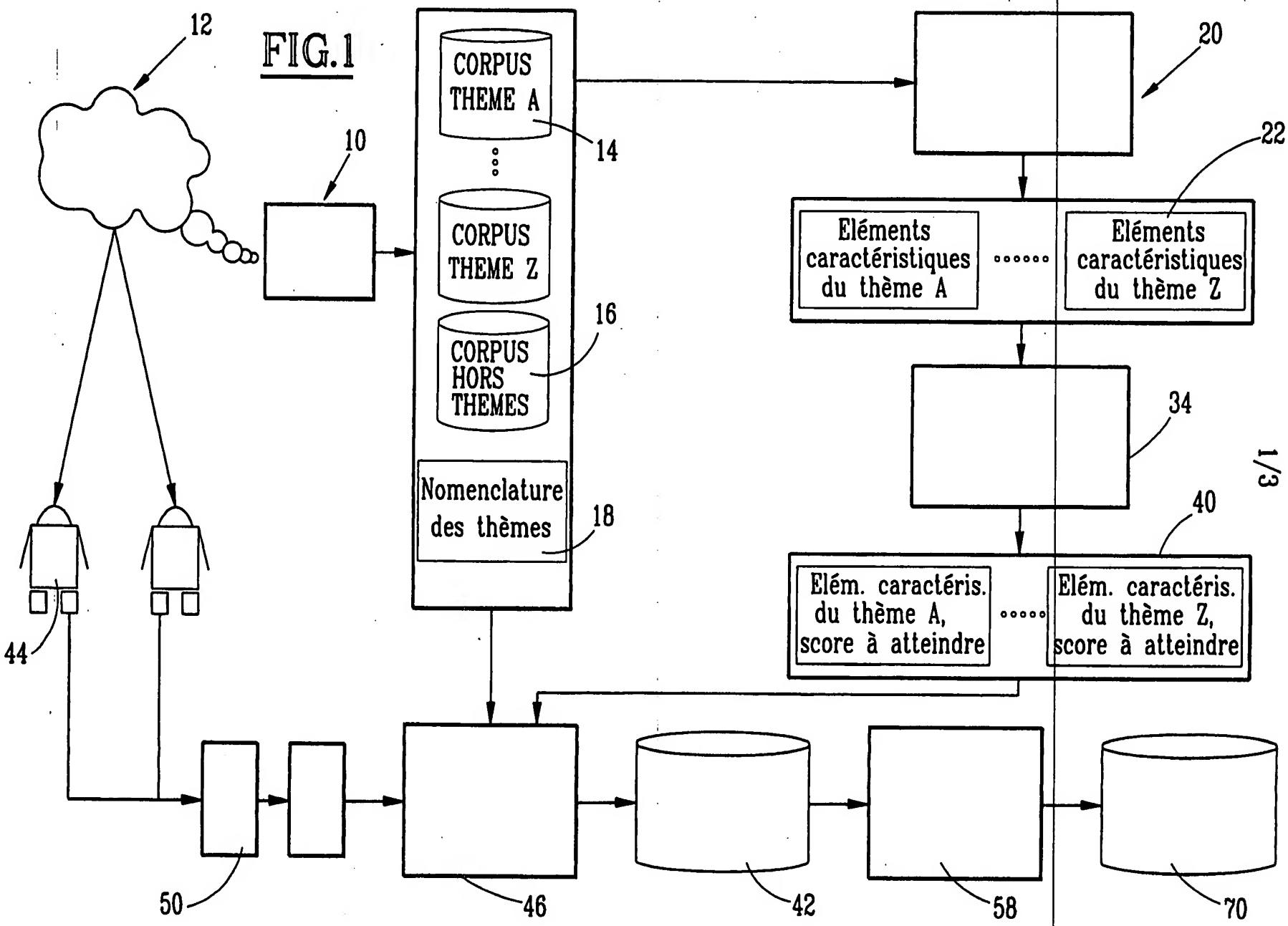
7. Procédé selon la revendication 5, caractérisé en ce que la valeur de seuil est réglée manuellement.

8. Procédé selon l'une quelconque des revendications 1 à 7, caractérisé en ce que les étapes d'identification des éléments caractéristiques de chaque thème contenu dans un document (50) sont réalisées au moyen d'une table de hachage.

FRANCE TELECOM

1/3

FIG.1



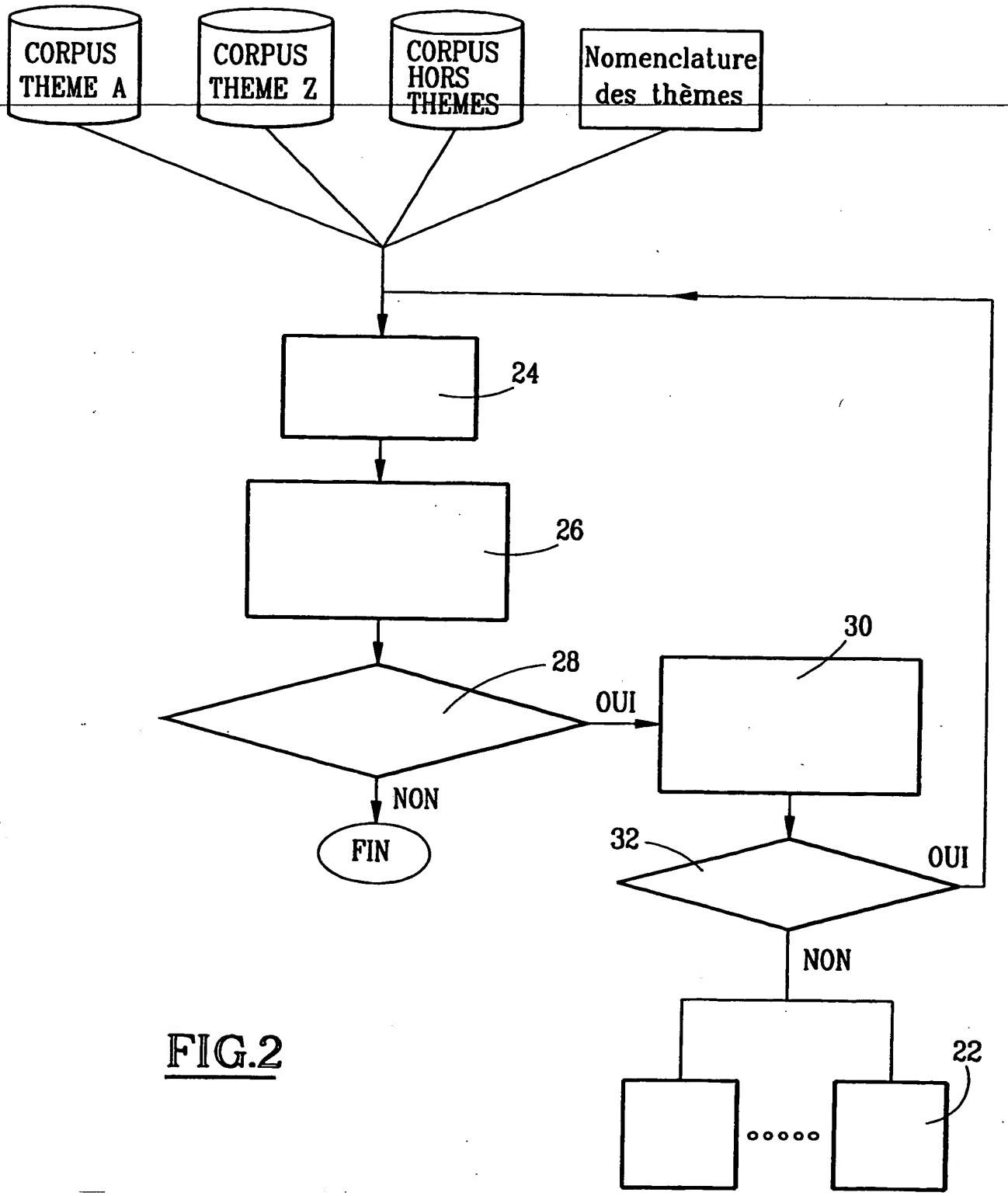

CORPUS THEME A
CORPUS THEME Z
CORPUS HORS THEMES
Nomenclature des thèmes
24
26
28
30
OUI
NON
FIN
32
OUI
NON
22

FIG.2

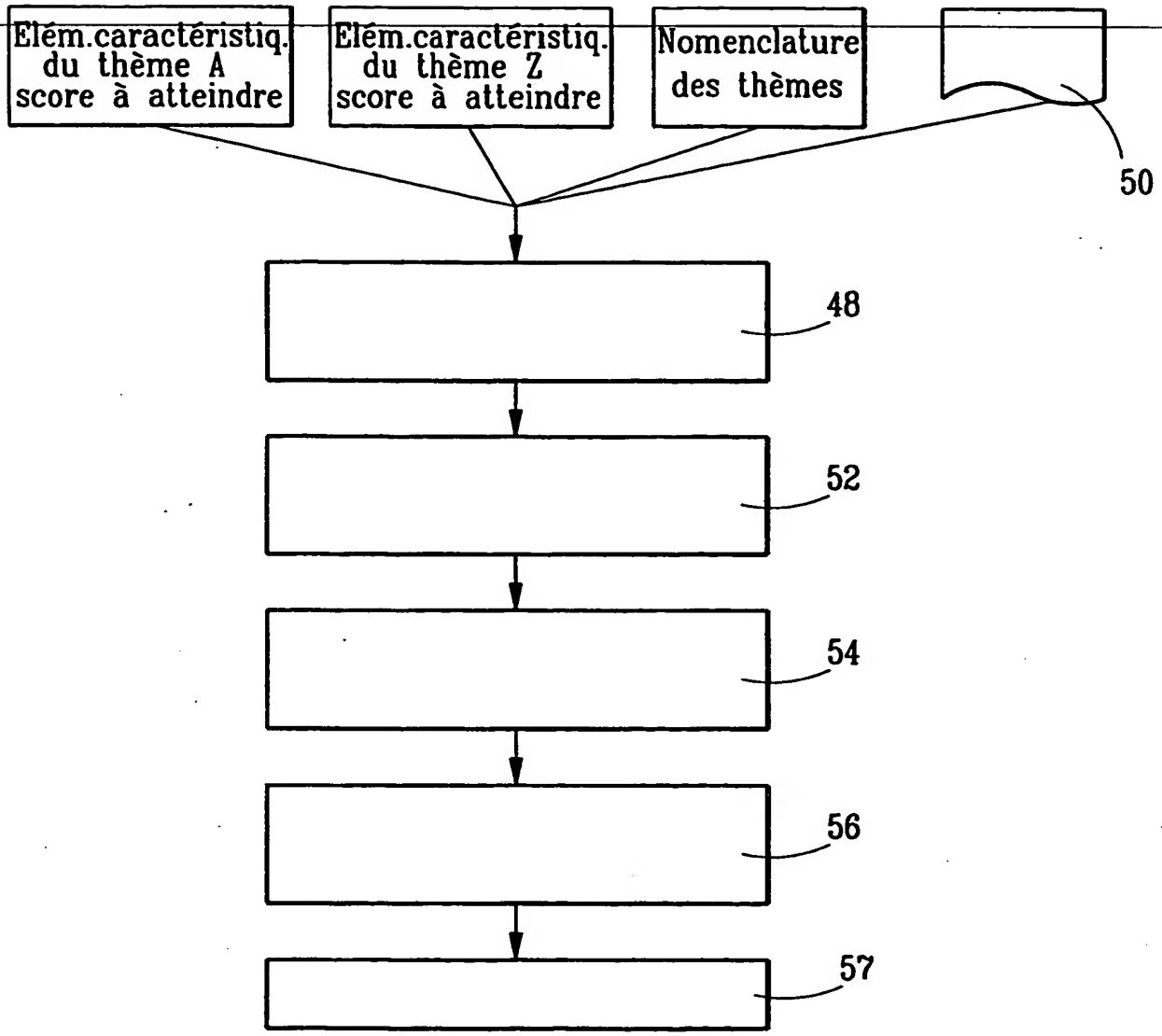

Elém.caractéristiq. du thème A score à atteindre
Elém.caractéristiq. du thème Z score à atteindre
Nomenclature des thèmes
50
48
52
54
56
57

FIG.3